Supplement for Individual Differences and Fitting Methods for the Two-Choice Diffusion Model of Decision Making

Roger Ratcliff and Russ Childers

The Ohio State University

This supplement adds information to the article. The topics are in an order that is consistent with the sections in the article.

Refinements to the Chi-Square Fitting Method

Over the last several years, we have added four refinements to chi-square fitting method.

1. When the data are divided into 5 quantiles, if the number of observations in a quantile is less than 7 or greater than 1, then a median split is used to form two bins, each with probability .5. When the data are divided into 9 quantiles, the median split for two bins is used when the number of observations is less than 10 (but see qualifications in point 2 below). Numbers of observations that are low occur mostly for errors in high accuracy conditions in which there are few errors. If the number of observations is 1, we compute a single value of chi-square ($(O-E)^2/E$, where O is the observed number of observations, 1 in this case, and E is the expected value from the model) for that condition and add this to the sum of the rest of the chi-square values for all the conditions and quantiles.

In the DMAT program, if the number of observations (usually errors) is less than 11, then error quantiles are not used in fitting, and drift rates are estimated poorly. DMAT provides a warning when this happens and indicates the fitted value may not be valid. For the other methods, all individual RTs are used, that is, RTs are not binned.

2. When the number of observations in a quantile is less than 7 for 5 quantiles or less than 10 for 9 quantiles and the median is outside the range of the .3 to .7 quantiles for correct responses, then we do not use the median. Instead, we combine the bins into a single bin to produce a single

value of chi-square as in point 1 (which of course does not use any information about RTs). We do this because when there are low numbers of errors in conditions with high accuracy, occasionally some of the error RTs can be spurious and (we assume) not from the decision process used in performing the task.

In contrast to binned methods, one extreme long or short RT can produce quite large biases in parameter estimates for methods that use every RT to produce a likelihood. As we show later, the fast-dm method with the KS statistic is robust to such short outliers.

3. Occasionally when there are few error responses, two consecutive error quantile RTs can be the same (e.g., 556 and 556 ms, with RTs with 1 ms resolution) and so the chi-square computation fails because the denominator of the chi-square is zero (because there will be zero probability mass between the two equal quantiles). Also, if quantile RTs are only 1 or 2 ms apart, the chi-square computation produces biased results because the probability mass between the quantiles is small. To address these problems, we added jitter to the raw RTs in the simulation studies described below by adding a uniformly distributed random number between -4 ms and +4 ms to each RT. This eliminated most of the problems that occurred when RT quantiles were too close together.

It might be thought that it is unrealistic to assume RTs are measured accurately to 1 ms. Generally, keyboards are polled and there can be systematic steps of 16 or 33 ms between successive polls of any key. We have measured delays up to 64 ms on older keyboards and found no problems because the variability in the keyboard times is combined (convolved) with the RT and the effect of the granularity in the keyboard response is very small. For example, if the SD in RT for a subject was 150 ms and the time between successive reads of a key was 32 ms, the SD in the keyboard would be 32/sqrt(12) ms (assuming a uniform distribution) so the SD for the combination (the square root of the sum of squares, i.e., sqrt(1502+9.22)=150.3) would be 150.3 ms. Therefore, variability in the keyboard response times is not an issue as long as the keyboard SD is lower than the SD in RTs.

4. Sometimes, accommodating very slow (perhaps spurious) error responses leads to estimates of across-trial variability in drift rate larger than they should be. Thus, because accuracy

values have to be fit, drift rates will be estimated to be larger. This means that drift rates covary with across-trial variability in drift rate parameters (e.g., Ratcliff & Tuerlinckx, 2002, Figure 6). An analog to this would be: If the SD in signal detection theory were increased, then to get the same hit rate, for example, the mean would have to be increased to compensate. In practice, with low numbers of observations in human subject data, sometimes long error RTs (some of which may be outliers) can cause the fitting program to produce values of across-trial variability in drift rate that are extremely large, and in compensation, this leads to large values of drift rate. Unless this is addressed, this can lead to drift rates several times larger than the values typically found with large numbers of observations (e.g., drift rates in the range 1.0 to 2.0 when the values should be in the range of .3 to .4).

This problem can be limited by placing upper and lower bounds on across-trial variability in drift rates (e.g., 0.08 and 0.3). The upper bound might be determined by examining the ranges of the variability parameter values from similar experiments with larger numbers of observations. Then upper and lower bounds can be set that are a little larger than the largest value from the subjects in the larger experiment. A lower bound can be set that is a little smaller than the smallest individual value. For fits for the simulation studies using the chi-square method with low numbers of observations, the value of the across trial variability in drift parameter was often estimated to be at the upper or lower bound for our programs that implemented these limits.

Simulation Study 1: Lexical Decision Design

The results of this study are generally the same as for the numerosity design. The major exception is that HDDM largely outperformed the other methods for this lexical decision design. Correlations between parameters used to generate simulated data and recovered parameters from fitting are shown in Table S1.

For most of the methods, recovered drift rates for the high-frequency words had high values, even with the largest number of observations (200) per condition and contaminants, because there were often no errors. As a result, the recovered values were more variable than for the drift rates for low- and very low-frequency words and nonwords and this led to correlations lower than for the drift rates for the other conditions. In the discussion that follows, we exclude

drift rates for [v1], the high-frequency word condition.

With 200 and 600 observations per condition and no contaminant RTs, all of the methods produced parameter values that were highly correlated with the generating values, above .75. Some of the drift rate correlations were lower than for the numerosity simulations because they used 1000 observations per condition (as opposed to either 200 or 600 used here). With 20 and 60 observations, HDDM, fast-dm, MLH, and the two chi-square methods produced correlations above .5. DMAT did considerably worse, with correlations dropping to the .1-.3 range. Correlations for the EZ were as low as .32.

With 4% contaminant RTs, for 200 and 600 observations per condition, HDDM, fast-dm, DMAT without contaminant correction, MLH, the two chi-square methods, and the EZ method all produced correlations greater than .72. DMAT with correction for contaminants gave lower correlations. The EZ method performed reasonably well, with correlations above .76.

With 4% contaminant RTs and 20 and 60 observations per condition, HDDM, fast-dm, the two chi-square methods, and the MLH method produced correlations greater than .53. The EZ method's correlations were as low as .26 and the two DMAT methods' were as low as .14.

Simulation Study 2: Lexical Decision Design

Generally, the results mirror those from the numerosity design and so provide a replication of its findings. Figures S1 and S2 show the recovered values for the 32 means and SD's for each of the parameters.

For the large numbers of observations (Figure S1), the chi-square methods show a little more bias for some drift rates, especially for the 9 quantile method, compared with the numerosity design. Fast-dm shows biases about the same size as for the numerosity design. HDDM as for the numerosity design shows some conditions with extreme biases that occur with the lower value (0.1) of boundary separation, and with 4% contaminants. The unbiased values from HDDM are closer to the true values with smaller SD's than for the chi-square methods. DMAT without contaminant correction shows larger SD's in model parameters than the other methods and shows some quite large biases in drift rates. DMAT with contaminant correction show even larger SD's and biases, values so large that 2 SD's for some drift rates include zero. The EZ method, as for the

numerosity design, shows quite small SD's, especially in drift rate, but very large biases, so large that some of the 0.1 drift rates are estimated to be zero. Similarly, nondecision time varies from less than 200 ms to 500 ms when the value used to generate the simulated data was 400 ms. This is expected because the EZ method assumes the starting point is halfway between the boundaries, and in a number of the sets of parameter values, the starting point is biased with values [a]/3 or 2[a]/3.

Thus, for this design, only the chi-square methods and fast-dm produce acceptable parameter recovery with small bias (relative to the other methods) and with small SD's in recovered parameters. If the spurious values produced by HDDM were fixed, then it would be acceptable also.

For the small numbers of observations (Figure S2), the results mirror those for the numerosity design with 40 observations per condition. The SD's in drift rates are a little larger but the biases are of similar size. The chi-square, fast-dm, and HDDM methods produce acceptable parameter recovery, with the lowest biases and SD's. This is somewhat of a puzzle as for the numerosity design because biases were less than with larger numbers of observations per condition. Finally, the EZ method and DMAT produce large biases and large SD's relative to the other methods.

DMAT warning messages

There are several warning messages provided by DMAT that indicate problems with parameter estimation. "The last convergence point was still a suspect result" indicates some of the across-trial variability parameter estimates may be wrong, "Hessian is not positive definite"/ "Hessian is not of full rank" indicates some parameter is not sufficiently identified by the data, and "Matrix is close to singular .." indicates some of the standard error estimates are likely biased. For the numerosity design with 40, 100, and 1000 observations per condition, the percentage of these three error messages across the 16 set of parameters and 64 simulated data sets per parameter are: 52%, 52%, and 13% for the three messages for N=40 per conditions, 23%, 31%, and 13% for the N=100 per conditions, and 2%, 6%, and 0% for the N=1000 per conditions. For the lexical design, the percentages were: 70%, 57%, and 18% for the smaller number of observations and 36%, 12%,

and 2% for the larger number of observations. These error messages indicate that the fits might be invalid (as spelled out in Vandekerckhove & Tuerlinckx, 2008). The results show that the DMAT package is warning that the fits are potentially flawed even with N=100 per condition, but are mainly good with N=1000 per condition.

EZ variants

We evaluated two other EZ variants, EZ2 and robust EZ, and found that neither was competitive with any of the methods examined so far. These variants were designed to address the problems with the assumption that the starting point is midway between the boundaries (EZ2, Grasman, Wagenmakers, & van der Maas, 2009) and the problem with contaminants (robust EZ, Wagenmakers, van der Maas, Dolan, & Grasman, 2008).

We examined these methods using just a few conditions because these are sufficient. For EZ2, we took 8 conditions from the lexical decision design with [z] at values 2[a]/3 and [a]/3 and with both the larger and smaller numbers of observations. For robust EZ, we used all 16 sets of parameter values with both 0% and 4% contaminants for the numerosity design for all three sets of numbers of observations. We found that the methods were rather poor at recovering parameter values and so did not pursue them further. We now describe the results.

Like the EZ method, EZ2 assumes no across-trial variability in model parameters. With both the lower and larger numbers of observations, parameter recovery was not good. For [a]=0.1, the average recovered value was 0.13 and for [a]=0.2, it was 0.21. For the narrow boundary separation (0.1), when [z]/[a] was 0.3, one drift rate was estimated to be 0 and the other 0.23 while when [z]/[a] was 0.7, one drift rate was estimated to be 0.23 and the other -0.3 when both were 0.2. Even in the cases in which [z]/[a] was 0.6, the mean drift rate was estimated to be 0.114 instead of 0.2. These results showed large deviations in conditions that the model was supposed to address and so we did not pursue the model further.

Robust EZ was developed to estimate the distribution of contaminants and then remove them from the analyses, in effect running EZ on the decontaminated distribution. The method involves first fitting a mixture distribution to the data: an exGaussian for the diffusion model and a uniform for contaminants, and then using the exGaussian distribution to compute the mean and

variance of the decontaminated distribution which are then used by the EZ method to produce diffusion model parameters.

We ran this model on the data set for the numerosity task in the second simulation (for the three sets of numbers of observations and 16 sets of parameter values). The results showed that, for example, when the original boundary separation was 0.2, the robust EZ analysis produced a value of about 0.138 for all the conditions with [a]=0.2. The proportion of outliers estimated did not track the actual number. When there were 40, 100, and 1000 observations per condition, we found the estimates of the proportion of contaminants were 38.4%, 20.3%, and 6.1% respectively (even though half the sets of parameter values had 0% contaminants and half 4%). Other model parameters were not as poorly estimated, for example, when [a] was 0.1, the estimated parameter was 0.088, [Ter] was 0.36 instead of 0.4, and the two drift rates were 0.078 and 0.22 instead of 0.1 and 0.2. But the problem with the wider boundary separation and proportion of contaminants made this variant uncompetitive.

Simulation Study 3: A Hierarchical Bayesian Diffusion Model

Normal Distributions Across Subjects

The means and SD's in the parameter values were those in the fourth and third lines from the bottom of Table 1 for which individual subjects varied from each mean according to a normal distribution.

Figure S3 shows the recovered values of the two drift rates, boundary separation, and nondecision time for the two methods for 40 observations per condition and 1000 observations per condition for the two methods (we used the results for the 100 observations condition later, but because the figures look the same as the ones presented, they are not shown). We present the plots in a different way from the other plots to highlight the issue of shrinkage of the parameters (regression to the mean).

Plotted on the x-axis are the values of the parameters from which the simulated data were generated. Plotted on the y-axis are the recovered values minus the value used to generate the simulated data (i.e., the residuals). If there is no shrinkage, the plots will be horizontal, but if there is shrinkage, the lines will have negative slope. Figures S3, S4, and 6 plot the results from the

hierarchical method and from the chi-square method, but if we plotted residuals, the functions would largely lie on top of each other and they would be hard to compare. To spread the values apart, they were offset by adding a constant for the chi-square method (the circles in the figure) and subtracting a constant for the hierarchical method (the x's in the figure). Any other biases in the recovered values will be seen as systematic deviations between the points and the horizontal solid lines. Also, the spread in the recovered parameter values, the residuals, can be used to compute the SD in the difference between the recovered parameter values and those used to generate the data. Thus, the vertical spread about the horizontal line visually represents this variability.

The top solid horizontal line in the figures is the mean offset for the chi-square method, the bottom solid line is the mean offset for the hierarchical method, and the dotted lines are regression lines. The correlations between the recovered values of the parameters (not the residuals) and the values used to generate the simulated data are shown in the headings of the panel.

For 40 observations per condition, correlations for the hierarchical method were higher than for the chi-square method for boundary separation and drift rates, but lower or the same for nondecision time. For 1000 observations per condition, the correlations for the chi-square method are higher for all the parameters.

The hierarchical method (but not the chi-square method) consistently underestimates nondecision time, with most of the points falling below the horizontal line. The hierarchical method also overestimates boundary separation with 4% contaminants (as does the chi-square method to a lesser degree), but does not do so with no contaminants.

For drift rates, there is moderate shrinkage in HDDM's recovered parameters because the dotted regression line for all the drift rates for all the conditions has negative slope. The drift rates for the chi-square method do not show systematic shrinkage. Perhaps surprisingly, there is no shrinkage in the boundary separation and nondecision time parameters for the HDDM model.

We discuss possible reasons in the main body of the paper for these results and the number of observations at which the chi-square method begins to outperform the hierarchical method, i.e., the cross-over point after the next two studies.

Uniform Distributions Across Subjects

For this study, the distributions for drift rates, boundary separation, and nondecision time came from uniform distributions with the same means and SD's as for the simulation just discussed. This means that each of these distributions across subjects is modestly mis-specified relative to assumed parameter distributions in the hierarchical model.

Figure S4 shows the results in the same way as for Figure S4 with normally distributed parameter values and results showed patterns that were qualitatively similar. The only major difference was that the shrinkage in drift rates was reduced relative to normal distributions in Figure S3, especially for the lower drift rates.

Very Low Numbers of Observations per Subject

Joachim Vandekerckhove suggested running simulations with the hierarchical method with the number of observations equal to the number of parameters plus 1. We used the numerosity design and 10 observations, 5 for each of the easy and hard conditions and the number of parameters for each subject was 9.

For each condition, we fit the hierarchical method to the first five trials in each condition from the data sets with 1000 observations per condition with the three distributions of parameter values from Figures S3, S4, and 6. Five observations per condition would be impossible to fit with the non-hierarchical methods because the low numbers of observations would produce estimates with very high variability and there would not be enough observations to produce meaningful RT distributions for both correct and error responses. Also, it is likely that the fitting programs would not converge on a solution most of the time.

The hierarchical fits for the three distributions of subject parameters from the prior simulations for 0% contaminants are shown in Figure S5 and the plots are in the same format as for Figures S3, S4, and 6. The correlations in parentheses are from the condition with 4% contaminants (the plots for 4% contaminants are very similar to those presented in Figure S5).

The first thing to note is that there is extreme shrinkage in the drift rates, especially for the normal and uniform distributions of parameters across subjects. In the plots, if the recovered parameters match those used to generate the data, the plot will be horizontal. Instead, the plots are diagonal for drift rates with a slope close to -1 which means that the drift rates are estimated to be

almost the same across conditions. For the most extreme case, for the uniform distribution of parameters for [v1] (the higher drift rate), the mean value of the drift rate parameters used to generate the simulated data is 0.284 and the mean recovered value is 0.247, but the SD in the drift rate parameters used to generate the simulated data is 0.094 and the SD in the recovered parameter values is 0.0054, i.e., only 6% of the variability used to generate parameters is recovered. For [Ter] for the two-population plot, there appears to be considerable shrinkage and the SDs in recovered values are only about half that of the values used to generate the simulated data.

One explanation for the extreme shrinkage in drift rates is in terms of what features of the data determine drift rates. Drift rates are most related to accuracy rates, and because there are only 5 observations per condition, there are not enough observations to constrain drift rates. Thus the hierarchical model constrains them to be nearly the same. In contrast, even with 5 observations, the RT differences from differences in boundary separation and nondecision time across subjects are large enough to produce moderate to large differences in recovered values of these parameters.

Despite the shrinkage, the correlations between the parameters recovered by HDDM and the parameters used to generate the simulated data are quite high. This is especially the case when the model is misspecified with two populations (leading to a larger spread of values) rather than one normal population. The correlations between the recovered parameter values and ones used to generate the data were high for boundary separation (.66 to .86), moderately high for nondecision time (.50 to .61) and even moderate for the drift rates (.35 to .61). Thus one could use the recovered parameter values to examine individual differences within the group. But because of the amount of shrinkage, the recovered parameter values could not be used to compare groups unless a study was done to examine biases for the different sets of parameter values for the groups and individual differences within groups.

Even though the hierarchical method with very few observations might recover individual differences relatively well, there are serious limitations with experiments based on this few observations. These include warmup and practice effects. For example, when undergraduate subjects begin an experiment, it is usual in our laboratory to ignore the first block of trials. If we are testing with a limited number of materials, we may test them for a few minutes on a related task, such as lexical decision or a perceptual task in order to familiarize them with the response

requirements. In the first few trials, subjects may be working out which fingers to use, still talking to the experimenter and so on. For older adults or adults who might have deficits, the warm up period might be significantly longer. We often train older adults, for example, for a few minutes on a different task to familiarize them with the stimulus presentation and response recording methods. An extreme case of warm up effects were experiments involving speed-accuracy instructions: Older adults required two or three full 45 min. experimental sessions before their performance was stable (e.g., Ratcliff et al., 2001, 2003; Thapar et al., 2003). Therefore, even though the hierarchical model may recover parameters reasonably well with very few observations, it would be a mistake to assume that the data were of adequate quality if this were the first reaction time task the subjects had encountered. Note that this is a problem with data and applies to all methods.

Benefits of Diffusion Model Applications: Low Numbers of Observation Designs

A problem that comes up for some experimental designs is what we call the "small-n" problem, which is that the number of observations in some conditions of an experiment is less than the number usually needed to use the model, for example, less than 100. In some cases (e.g., neuropsychological testing and clinical applications), this is because only a limited amount of time is available for testing. In other cases, the limit is the number of items that are available, the number that can be constructed, or the number that can be used in an experiment before subjects develop expectations about what kind of test items to expect. As an example, White, Ratcliff, Vasey, and McKoon (2009) investigated differences between mildly depressed (dysphoric) subjects and control subjects in a lexical decision experiment. In the depression literature, there are only about 30 words for which there is good agreement among researchers that they express the negative meanings that are relevant to dysphoric individuals. Experiments on reading comprehension provide other examples. McKoon and Ratcliff (1986, 1992, 2013) tested whether subjects comprehend inferences like "the actress died" from sentences like "The director and the cameraman were ready to shoot close-ups when the actress fell from the 14th story roof" in an item recognition experiment. Constructing items like this is difficult and the number of items needs to be kept small so that subjects do not begin making the inferences on the basis of explicit strategies; for example, McKoon and Ratcliff's 2013 experiment had only 32 such items.

Our solution to the small-n problem is to include many filler items in an experiment and use the RTs and accuracy for these items to constrain fitting the model to the items in the experimental conditions. In White et al's (2009) experiment, there were 540 filler words and 510 nonwords and in McKoon and Ratcliff's (2013), there were 416 words that had been studied and 416 words that had not. The model is fit simultaneously to all the data with the result that the filler items largely determine all the parameters of the model except the drift rates for the conditions of interest. These drift rates can be estimated with enough precision to compare performance in conditions to each other and to compare performance of one subject group to another (for example, doubling of power, see White, Ratcliff, Vasey, & McKoon, 2010).

Goodness of Fit

Ratcliff, Thapar, Gomez, and McKoon (2004) examined the effect of moving a .1 probability mass from one quantile bin (so the .2 probability mass became .1) to another adjacent quantile bin (so the .2 probability mass became .3), i.e., the first constraint in the prior paragraph. They found that the increment to the chi-square was 0.133*N, so for N=100, the increment would be 13.3 and for N=1000, the increment would be 133. Thus, for example, for two conditions and 5 quantiles with 13 degrees of freedom (22 from data minus 9 model parameters), the change in chi-square would be over half the critical value (of 22.4) for N=100 and it would be five times the critical value of N=1000. These increments mean that even relatively small systematic misses in the proportions are accompanied by quite large increases in the chi-square. Furthermore, the chi-square statistic is a very conservative statistic so that even small systematic differences between theory and data show large increases in the chi-square as the number of observations increases.

Combining Parameters

Because there are correlations between model parameters for the fits, we attempted to see if combinations of them provided a more compact description of the results. We ran exploratory factor analysis on the model parameters (for both low and high numbers of observations in the numerosity design) to see if some combination of parameter values produced better correlations between fitted parameters and those used to generate the simulated data. Although some of the factors made sense, the loadings were weak and did not support the attempt to extract factors from

combinations of the model parameters (for these experiments).

An Example of Comparing Variability in Subjects and Parameter Estimates

To provide a concrete example of the effect of variability in parameter estimates relative to individual differences, 1000 normally distributed random numbers were generated with SD *s* to produce values *x*. From these, three sets of random numbers (*u*, *v*, and *w*) were generated with means *x* and SD's either *s*, .75*s*, or .5*s*. These represent examples in which the *x* values represent individual differences, and the *u*, *v*, and *w* values represent values with SD's in the estimates of *x* of either *s*, .75*s*, or .5*s*. The correlations between *x* and *u*, *x* and *v*, and *x* and *w* were .72, .80, and .89 which shows that even if the SD in estimation is the same value as the SD in individual differences (*s*), the ceiling on the correlation would be reduced from 1 to only .72.

Supplementary Figure Captions

Figure S1. Plots of boundary separation, nondecision time, and three drift rates for the lexical decision design with 300, 300, and 600 observations for the high and low frequency words and nonwords respectively for Simulation Study 2. Each row shows a different fitting method. On the x-axis is plotted the mean of the parameter values and on the y-axis, in a horizontal row at the value of the parameter used to generate the simulated data is plotted 1 SD error bars. The thin vertical line represents the values used to generate the simulated data. Movement away from the vertical line on the x-axis represents bias in the recovered parameter values and a large spread of the error bars represents high variability in the recovered parameter values. There are two values of boundary separation and two values of drift rate, hence the two vertical lines and the vertical separation of the points.

Figure S2. The same plot as in Figure S1, but for the lexical decision design with 30, 30, and 60 observations for the high and low frequency words and nonwords respectively for Simulation Study 2. As for Figure 5, the DMAT results had SD's that overlapped zero and so are not shown.

Figure S3. Plots of the recovered values of parameters from the hierarchical fitting method and for the chi-square method for the numeracy design with normally distributed population parameters for 40 and 100 observations per condition and for 0 and 4% contaminants. Plotted on the y-axis are the recovered parameter values minus the values used to generate the simulated data (i.e., the residuals) offset by the amount represented by the thick horizontal lines. The circles are for the chi-square method and the crosses are for the hierarchical method. The dashed lines are regression lines for the two methods (shrinkage of model parameters results in a negative slope). The numbers at the top of each plot are the correlations between the recovered values and those used to generate the simulated data for the two methods.

Figure S4. Plots of the recovered values of parameters from the hierarchical fitting method and for the chi-square method for the numeracy design with uniformly distributed parameters. Other details of the plots are the same as for Figure S3.

Figure S5. Plots of the recovered values of parameters from the hierarchical fitting method

for the numeracy design with normally distributed population parameters for 5 observations per condition. The plots are otherwise the same as for Figure S3 (but there are only fits for the HDDM method).

Figure S6. Plots of boundary separation, nondecision time, and two drift rates for the numeracy design with 1000 observations per condition in Simulation Study 2 as in Figure 5. The top row shows the original HDDM parameter recovery and next two show the effects of fixing the proportion of contaminants; this reduced the size and number of spurious results.